ANALYSIS AND MODELLING OF EARLY STAGE DIABETES RISK PREDICTION DATASET

- M.Shozab Hussain 23100174
- Ahmed Tahir Shekhani 23100197
- Jahanzeb Raza 23100003
- Syed Talal 23100176
- Ali Asghar 23100198

DESCRIBING THE DATASET

-> The dataset contains a total of **8,857** observations and **17** features. This dataset contains reports of diabetes-related symptoms of **521** persons. It includes data about people including symptoms that may cause diabetes.

-> Each row in the dataset represents the record of a single patient

-> Features are : **Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity, Class**

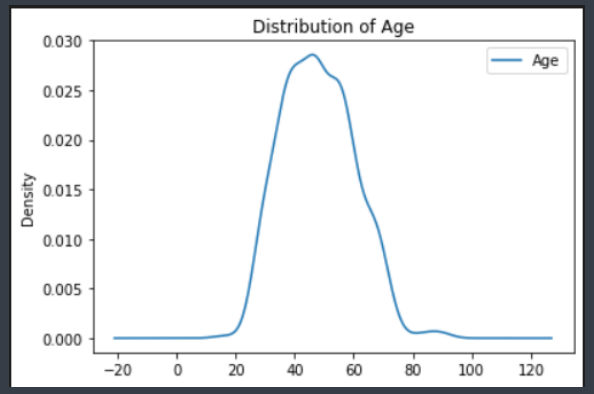
Dataset Cleaning

- -> Missing Values
- -> Binary Values
- -> Column names
- -> Age groups

STATISTICAL INFERENCE AND DATA VISUALIZATION

- Spread of Patients in Each Age Group
- Number of Patients with each Symptom
- Correlation Matrix to Visualize Relationship of Features
- Conclusions from Visualization:

Drop itching, Delayed Healing and Age Group



MACHINE LEARNING MODEL

- Preparing feature set and test label
- Splitting into data test and train data
- Logistic Regression
- 5-fold cross validation
- Predicting test labels

```
The Feature Set of our Model:
 [-0.66,  1.  ,  0.  , ...,  1.  ,  1.  ,  1.  ],
 [ 0.82,  1.  ,  0.  , ...,  0.  ,  1.  ,  0.  ],
 [-0.58,  1.  ,  1.  , ...,  1.  ,  1.  ,  0.  ],
 ...,
 [ 0.82,  0.  ,  1.  , ...,  1.  ,  0.  ,  1.  ],
 [-1.32,  0.  ,  0.  , ...,  0.  ,  1.  ,  0.  ],
 [-0.5 ,  1.  ,  0.  , ...,  0.  ,  0.  ,  0.  ]
```

```
Train set: (364, 16) (364, 1)
Test set: (156, 16) (156, 1)
```

MODEL RESULTS

Accuracy:

$$\frac{TP+TN}{TP+TN+FP+FN}$$

Recall or Sensitivity:

$$\frac{TP}{TP+FN}$$

Precision:

$$\frac{TP}{TP+FP}$$

F1 Score:

$$\frac{2*precision*recall}{precision+recall}$$

TP = True positive
TN = True Negative
FP = False Positive
FN = False Negative

Our Dataset result:

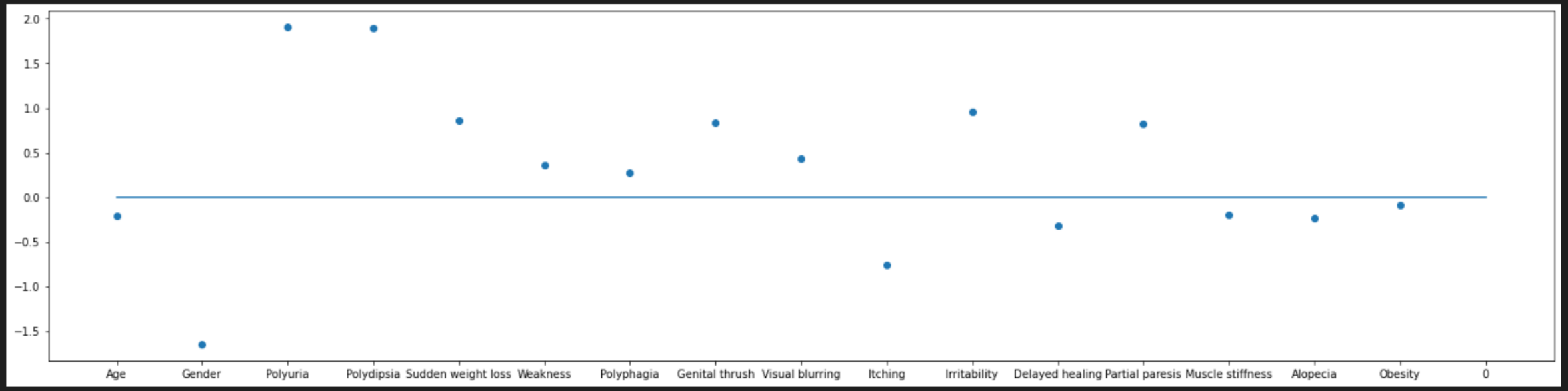
Accuracy_train	F1_score_train
0.925824	0.936471

Accuracy_test	F1_score_test
0.942308	0.955665

Precision_train	Recall_train	r2_score_train
0.961353	0.912844	0.691215

Precision_test	Recall_test	r2_score_test
0.960396	0.95098	0.745098

VISUALIZATION OF WEIGHTS



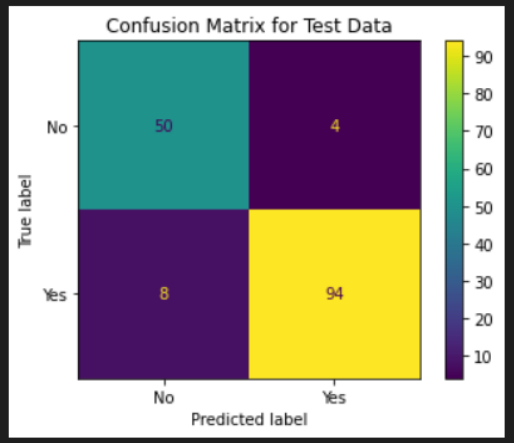
Weights closer to zero have very less effect on model results (predicted value of either having a diabetes or not)

COMPARING RESULTS BY DROPPING FEATURES

Dropped our proposed features

	Accuracy_train	F1_score_train	Precision_train	Recall_train	r2_score_train
0	0.914835	0.9274	0.947368	0.908257	0.645469

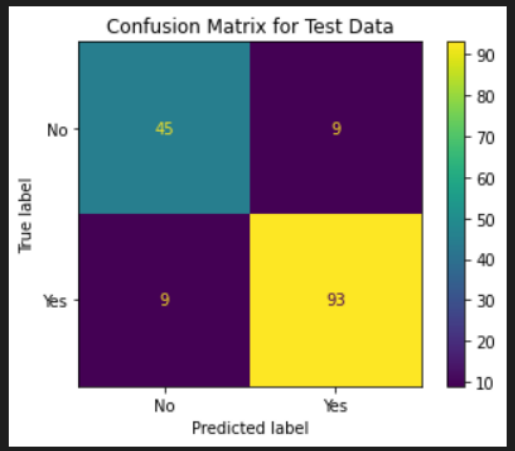
	Accuracy_test	F1_score_test	Precision_test	Recall_test	r2_score_test
0	0.923077	0.94	0.959184	0.921569	0.660131



Dropped Gender Column

	Accuracy_train	F1_score_train	Precision_train	Recall_train	r2_score_train
0	0.903846	0.916067	0.959799	0.876147	0.599724

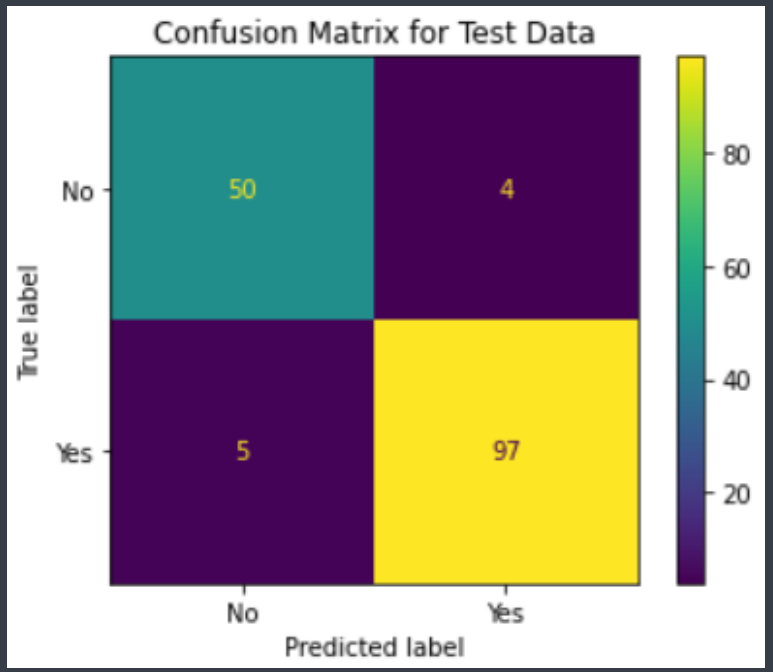
	Accuracy_test	F1_score_test	Precision_test	Recall_test	r2_score_test
0	0.884615	0.911765	0.911765	0.911765	0.490196



DID OUR MODEL SOLVE OUR PROBLEM AND ITS REAL-WORLD APPLICATIONS

- THE PROBLEM: CAN OUR MODEL CORRECTLY PREDICT WHETHER A PERSON HAS DIABETES, JUDGING BY THE SIGNS AND SYMPTOMS THEY ARE EXPERIENCING.
- THE OUTCOME: SIGNIFICANTLY SMALL FALSE NEGATIVE PERCENTAGE I.E., APPROX. 3% SUGGESTS THAT IT CAN.

$$\frac{5}{156} \times 100 \approx 3\%$$



Our model's real-world applications

- Model can be implemented from where the data was collected, i.e., Bangladesh.
- Similarly, our model will deliver promising results in countries with the same demographic (e.g., Pakistan, India, etc.).
- Quick results and minimal requirements offer cheap and quick diagnosis.
- Can be utilized in areas that have limited or no access to healthcare services (villages, tough terrain areas, etc.).